Claims

1. A method for assembly of a plurality of reads from a genomic region, the method comprising the steps of:

providing a plurality of reads from a genomic region;

indexing a plurality of read subsequences for each of the plurality of reads, each subsequence having an associated read with which it corresponds;

extracting from the indexed subsequences a plurality of read pairs that have a predetermined number of subsequences in common; and

merging the read pairs along a continuum.

2. The method of claim 1, comprising the step of providing a plurality of reads generated from sequencing both ends of a plurality of DNA segments, each read having associated linking information comprising an associated orientation relative to a read from an opposite end of the DNA segment, and an associated distance from the read on the opposite end of the DNA segment.

3. The method of claim 1, comprising the step of providing a plurality reads that are reverse complements of a plurality of reads provided by sequencing both ends of a plurality of DNA segments.

4. The method of claim 1, comprising the step of sorting the indexed read subsequences.

5. The method of claim 1, comprising the step of discarding read subsequences having more than a cutoff number of occurrences from the plurality of indexed subsequences.

6. The method of claim 1, comprising the step of indexing a plurality of read subsequences of a predetermined length for each of the plurality of reads.

7. The method of claim 6, wherein the predetermined length for each of the plurality of reads is between about 12 and about 32 bases long.

8. The method of claim 1, comprising the step of indexing a plurality of subsequences for each read, the index comprising for each subsequence an associated read and an associated position on the read with which it corresponds.

9. The method of claim 1, comprising the step of performing alignments on the plurality of read pairs having a predetermined number of subsequences in common.

10. The method of claim 8, comprising the steps of:

performing alignments on the plurality of read pairs having a predetermined number of subsequences in common; and

using the associated position on the reads with which the subsequences correspond to verify overlap.

11. The method of claim 2, comprising the step of using the linking information associated with the reads to confirm that the merged pairs are merged correctly.

12. The method of claim 2, comprising the step of using the associated linking information to an ambiguity in the merged reads.

13. The method of claim 12, comprising the step of identifying a repeat region and a set of unique regions.

14. The method of claim 13, comprising the step of linking pairs of unique regions using the linking information associated with the reads in the unique regions.

15. The method of claim 14, comprising the step of inserting the repeat region between each linked pair of unique regions with which the repeat region corresponds.

16. The method of claim 13, comprising the step of merging linked pairs of unique regions using the linking information associated with the reads in the unique regions.

17. A method for assembly of merged reads from a genomic region, the method comprising the steps of:

providing one or more sets of merged reads from a genomic region comprising a set of reads having associated linking information;

using the associated linking information to identify an ambiguity in the merged reads;

identifying a repeat region and a set of unique regions; and

linking pairs of unique regions using the linking information associated with the reads in the unique regions.

18. The method of claim 17, comprising the step of inserting the repeat region between each linked pair of unique regions with which the repeat region corresponds.

19. The method of claim 17, comprising the step of merging linked pairs of unique regions using the linking information associated with the reads in the unique regions.

20. An article of manufacture having computer-readable program means embodied thereon for assembly of a plurality of reads from a genomic region, the article comprising:

computer-readable program means for providing a plurality of reads from a genomic

region;

computer-readable program means for indexing a plurality of read subsequences for each of the plurality of reads, each subsequence having an associated read with which it corresponds;

computer-readable program means for extracting from the indexed subsequences a plurality of read pairs that have a predetermined number of subsequences in common; and

computer-readable program means for merging the read pairs along a continuum.

21. An article of manufacture having computer-readable program means embodied thereon for assembly of merged reads from a genomic region, the article comprising:

computer-readable program means for providing one or more sets of merged reads from a genomic region comprising a set of reads having associated linking information;

computer-readable program means for using the associated linking information to identify one or more ambiguities in the merged reads;

computer-readable program means for identifying a repeat region and a set of unique regions; and

computer-readable program means for linking pairs of unique regions using the linking information associated with the reads in the unique regions.